Class 03: Decision Theory

Martin Slawski

Volgenau School of Engineering
Department of Statistics

February 1, 2018

Decision Theory

Last class, we informally introduced a series of learning problems, including

- Classification
- Regression
- Clustering
- Density Estimation

In this class, we will introduce a framework within which these problems can be formalized.

Decision Theory

Recall that we have distinguished the **supervised** and **unsupervised** setting.

In the **supervised** setting, the dataset $\mathcal{D}$ of sample size n is from $(\mathcal{X} \times \mathcal{Y})^n$, where

- $\mathcal{X}$ is the domain of the *inputs*, *input variables*, *covariates*, *predictors*, *attributes* or *features* (all synonymous). Typically, $\mathcal{X} \subseteq \mathbb{R}^d$.
- $\mathcal{Y}$ is the domain of the *labels*, *outputs*, *responses*, *response variables*. $\mathcal{Y}$ can be binary, integer-valued, real valued, or more complex (structured output, e.g. graphs).

In the **unsupervised** setting, we do not observe any labels, i.e. $\mathcal{D}$ takes values in $\mathcal{X}^n$.

Decision Theory

Most learning problems that we shall discuss in this class amount to finding a function from some class $\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathbb{R}^T\}$ (most of the time, $T = 1$).

$\mathcal{F}$ is often referred to as **hypothesis class**. In many cases, $\mathcal{F}$ can be identified with a parameter set Θ as familiar from parametric statistical inference.

A learning problem is specified by choosing a **loss function**

$$L : \mathcal{Y} \times \mathbb{R}^T \rightarrow \mathbb{R}_+ \quad \text{(supervised)},$$
$$L : \mathcal{X} \times \mathbb{R}^T \rightarrow \mathbb{R}_+ \quad \text{(unsupervised)}.$$

to measure how well f accomplishes the goal of interest (classification, regression, etc.) for a single datum.

Decision Theory

In statistical machine learning, we assume that the data are i.i.d. samples from a probability distribution P on $\mathcal{X} \times \mathcal{Y}$ (respectively only on $\mathcal{X}$).

This leads to the notion of **risk** or **expected loss**

$$\begin{aligned} R(f) &= \mathbf{E}_{(X,Y) \sim P}[L(Y, f(X))] \quad \text{(supervised)}, \\ R(f) &= \mathbf{E}_{X \sim P}[L(X, f(X))] \quad \text{(unsupervised)}. \end{aligned}$$

The optimal risk is denoted by

$$R^* = \inf_{f \colon f \text{ measurable}} R(f).$$

The corresponding minimizer is denoted by f_L^*.

Decision Theory

Scenario I: binary classification

$\mathcal{Y} = \{-1, 1\}$, $0 - 1$ loss or misclassification error

$$L_{\text{0-1}}(y, f(x)) = I(y \neq \text{sign}(f(x))\,) = \begin{cases} 0 & \text{if } y = \text{sign}(f(x)), \\ 1 & \text{else.} \end{cases}$$

Here, any of the two possible ways of misclassification are treated equally. This is often not appropriate and motivates the following loss function:

$$\begin{aligned} L_{\text{0-1},C}(y, f(x)) = C_{1,-1} I(y = 1 \wedge \text{sign}(f(x)) = -1) + \\ + C_{-1,1} I(y = -1 \wedge \text{sign}(f(x)) = 1). \end{aligned}$$

Straightforward to generalize to multi-class classficiation.

Decision Theory

The corresponding risk is given by the probability of error or misclassification rate

$$R(f) = \mathbf{E}[I(Y \neq \mathsf{sign}(f(X))\,)] = \mathbf{P}(Y \neq \mathsf{sign}(f(X)))$$

Theorem

For the $0-1$ *loss, the risk is minimized by the* **Bayes rule** *(or* **Bayes classifier***)*

$$f^*(x) = \operatorname*{argmax}_{y \in \{-1,1\}} \mathbf{P}(Y = y | X = x), \; x \in \mathcal{X}.$$

The associated value of the risk $R^* = R(f^*)$ is called **Bayes risk**.

Decision Theory

Proof.

Using iterated expectation, we get

$$\mathbf{E}_{X,Y}[L_{\text{0-1}}(Y, f(X))] = \mathbf{E}_X\, \mathbf{E}_Y[L_{\text{0-1}}(Y, f(X))|X]$$

Now note that for any fixed $x \in \mathcal{X}$

$$\begin{aligned}\mathbf{E}_Y[L_{\text{0-1}}(Y, f(x))|X = x] &= \mathbf{P}(Y \neq \text{sign}(f(x))|X = x)\\ &= \begin{cases}\mathbf{P}(Y = 1|X = x) & \text{if sign}(f(x)) = -1,\\ \mathbf{P}(Y = -1|X = x) & \text{if sign}(f(x)) = 1.\end{cases}\end{aligned}$$

Hence, in order to minimize $\mathbf{P}(Y \neq \text{sign}(f(x))|X = x)$, we choose $f(x)$ as $f^*(x) = \text{argmax}_{y \in \{-1,1\}}\, \mathbf{P}(Y = y|X = x)$.

This reasoning holds for arbitrary $x \in \mathcal{X}$. □

Decision Theory

It is important to note that the Bayes classifier is not a practical scheme, because we do not know the underlying distribution on $\mathcal{X} \times \mathcal{Y}$.

Instead, the Bayes classifier and the Bayes risk serve as **benchmark** for practical schemes discussed in subsequent material.

Ideally, we would like to be (nearly) as good as the Bayes classifier as we see more and more samples.

The terminology "Bayes ..." has to do with the relation

$$\mathbf{P}(Y = y | X = x) = \frac{f_{X|Y=y}(x)}{f_X(x)} \mathbf{P}(Y = y), \quad x \in \mathcal{X}, \; y \in \mathcal{Y}.$$

Note that

$$\operatorname*{argmax}_{y \in \mathcal{Y}} \mathbf{P}(Y = y | X = x) = \operatorname*{argmax}_{y \in \mathcal{Y}} f_{X|Y=y}(x) \, \mathbf{P}(Y = y).$$

Decision Theory

In this context, the class proportions $\{\mathbf{P}(Y = y)\}_{y \in \mathcal{Y}}$ are called prior (or a priori) probabilities.

After observing x, we obtain posterior probabilities $\mathbf{P}(Y = y|X = x), \ y \in \mathcal{Y}$.

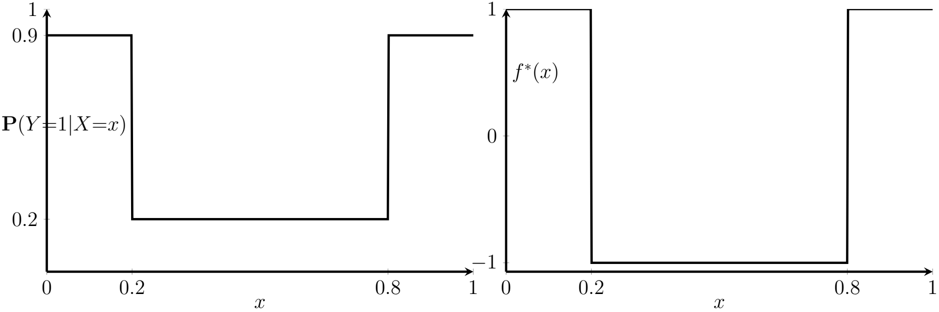
The standard example for X is a diagnostic medical test, with two possible outcomes (positive or negative test).

Small Bayes risk R^* does not necessarily mean that the classification problem is easy – think of highly unbalanced class proportions (e.g. Y represents a rare disease).

If $R^* = 0$ one typically says that the classes are *perfectly* separable.

Decision Theory

Determing the Bayes classifier and the Bayes risk in a toy example: $\mathcal{X} = [0, 1]$, marginal distribution P_X of X is the uniform distribution.



We compute $R^* = 2 \cdot (0.1 \cdot 0.2) + 0.2 \cdot 0.6 = 0.16.$

Decision Theory

Issue with the 0-1 loss: leads to an **NP**-hard optimization problem "in practice" (to be explained later).

$\rightsquigarrow$ convex **margin-based** loss funtions

Definition

For $f : \mathcal{X} \to \mathbb{R}$ and $y \in \{-1, 1\}$, we call $yf(x)$ the **margin** *of f at (x, y).*

The margin can be seen as a "score of agreement" between y and the prediction $f(x)$.

In the sequel, we consider a series of **margin-based** loss functions of the form

$$L(y, f(x)) = L(yf(x))$$

Decision Theory

The 0-1 loss can be written as margin-based loss

$$L_{\text{0-1}}(y, f(x)) = L_{\text{0-1}}(yf(x)) = I(yf(x) < 0).$$

We are interested in alternative loss functions having the following properties:

- L is a convex function of the margin,
- L upper bounds $L_{\text{0-1}}$

Decision Theory

1) Hinge loss:

$$L_{\text{hinge}}(y, f(x)) = \max\{0, 1 - yf(x)\}$$

2) Exponential loss:

$$L_{\text{exp}}(y, f(x)) = \exp(-yf(x))$$

3) Logistic loss:

$$L_{\text{logistic}}(y, f(x)) = \log_2(1 + \exp(-yf(x)))$$

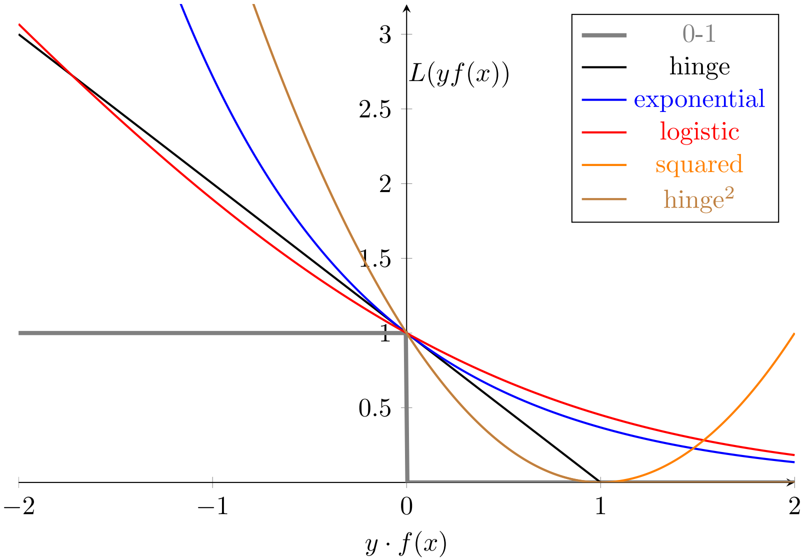
4) Squared loss: (not recommended)

$$L_{\text{squared}}(y, f(x)) = (1 - yf(x))^2$$

5) Squared hinge loss:

$$L_{\text{hinge}^2}(y, f(x)) = \max\{0, 1 - yf(x)\}^2$$

Decision Theory



Decision Theory

Comparison of loss functions:

The hinge loss is the **tightest convex upper bound** to the 0-1 loss. However, it is **non-smooth**.

Both the hinge loss and the logistic loss are 1-**Lipschitz**. They are hence more robust than the other losses.

The exponential loss and logistic loss are **smooth**. The squared hinge loss is only $1\times$ differentiable at 1.

All these loss functions satisfy a basic consistency requirement to be introduced below.

Decision Theory

Basic consistency requirement:

For a margin-based loss function L, write

$$f_L^*(x) = \operatorname*{argmin}_{\alpha} \mathbf{E}[L(Y\alpha)|X = x], \ \ x \in \mathcal{X}.$$

Definition (Bartlett, Jordan, McAuliffe, 2006)

A margin-based loss function L is called **classification-calibrated** if $\text{sign}(f_L^*(x)) = f^*(x), \ x \in \mathcal{X}$, where f^* is the Bayes classifier.

Intuition: by replacing $L_{\text{0-1}}$ with a classification-calibrated loss, we still manage to implement the Bayes classifier since the minimizer of this loss is equivalent to the Bayes classifier.

We often speak of a **surrogate** loss.

Decision Theory

Scenario II: Regression

$\mathcal{Y} = \mathbb{R}$.

For regression, the setup is less intricate as for classification which has a distinguished loss (0-1) that is replaced by a surrogate loss.

For regression, we typically consider **distance-based** losses of the form

$$L(y, f(x)) = L(y - f(x)).$$

For computational reasons, convex L is preferred. Again, we write

$$f_L^*(x) = \operatorname*{argmin}_{\alpha} \mathbf{E}[L(Y - \alpha) | X = x], \;\; x \in \mathcal{X}.$$

Decision Theory

1) Squared loss:

$$L_{\text{squared}}(y, f(x)) = (y - f(x))^2$$

---

2) Absolute loss:

$$L_1(y, f(x)) = |y - f(x)|$$

---

3) ε-insensitive loss

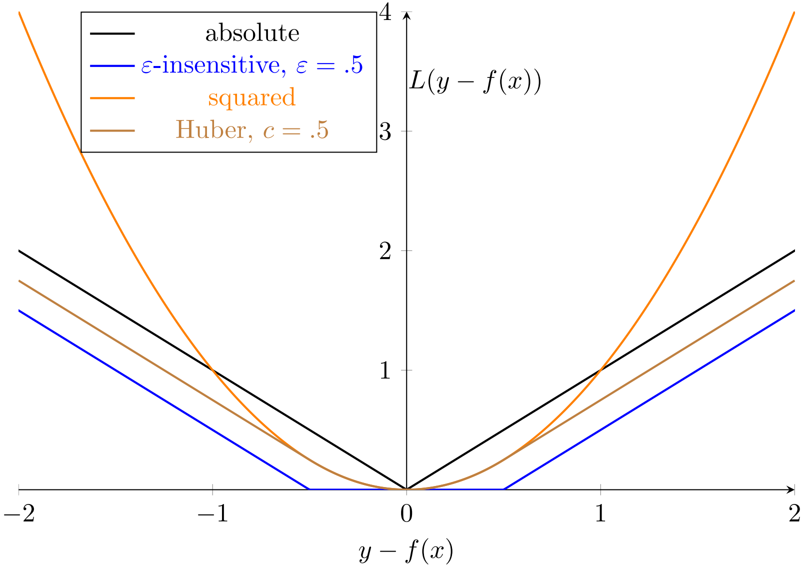
$$L_\varepsilon(y, f(x)) = |y - f(x)| I(|y - f(x)| \geq \varepsilon)$$

---

4) Huber loss

$$L_{\text{Huber}}(y, f(x)) = \begin{cases} \frac{1}{2c}(y - f(x))^2 & \text{if } |y - f(x)| \leq c \\ |y - f(x)| - c/2. & \text{if } |y - f(x)| > c. \end{cases}$$

Decision Theory



Decision Theory

Comparison of loss functions:

Squared loss and absolute correspond to Gaussian and Laplacian noise, respectively. The latter has heavier tails. Therefore, the absolute loss behaves much more robustly. This can also been seen from the observation that

$$\begin{aligned} f^*_{L_{\text{squared}}}(x) &= \mathbf{E}[Y|X=x], \\ f^*_{L_1}(x) &= \text{median}(Y|X=x). \end{aligned}$$

The L_1 loss is 1-Lipschitz, but non-smooth.

The Huber loss can be seen as a compromise between squared and absolute loss. It is differentiable.

Decision Theory

Scenario III: Clustering / Vector Quantization

This is an **unsupervised problem**, so there is only $\mathcal{X}$ and no $\mathcal{Y}$.

One possible loss function is the K-**means** objective:

$$L(x, f(x)) = \|x - f_{\mathcal{M}}(x)\|_2^2,$$

where, given $\mathcal{M} = \{\mu_1, \ldots, \mu_K\} \subset \mathbb{R}^d$,

$$f_{\mathcal{M}}(x) = \underset{\mu \in \mathcal{M}}{\operatorname{argmin}} \|x - \mu\|_2^2.$$

Note that the corresponding risk is given by

$$R(f_{\mathcal{M}}) = \mathbf{E}[\|X - f_{\mathcal{M}}(X)\|_2^2]$$

For $K = 1$, it is easy to see that

$$f_L^*(x) = \mathbf{E}[X], \;\; x \in \mathcal{X}.$$

Decision Theory

Scenario IV: Density Estimation

Suppose the distribution P_X on $\mathcal{X} \subseteq \mathbb{R}^d$ has a density f_X.

In Class 01, we have considered the loss

$$L(x, f(x)) = (f_X(x) - f(x))^2.$$

The corresponding risk

$$R(f) = \mathbf{E}[(f_X(X) - f(X))^2]$$

is the integrated mean squared error.

Decision Theory

Our goal is to minimize the risk (or synonymously, expected loss)

$$R(f) = \mathbf{E}_{(X,Y)\sim P}[L(Y, f(X))]$$

over "all possible f".

In practice, we are only given a sample $\mathcal{D}$ consisting of n i.i.d. pairs distributed according to P.

$\rightsquigarrow$ Replace the risk R by its empirical counterpart

$$R_{\textsf{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)).$$

By the law of large numbers, we know that for any **fixed** f

$$R_{\textsf{emp}}(f) \to R(f) \text{ in probability, as } n \to \infty.$$

Hence, we expect/hope that minimizing the empirical risk $R_{\textsf{emp}}$ is a good proxy for minimizing the risk R.

Decision Theory

$$R_{\mathsf{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)).$$

Minimizing the empirical risk is called **empirical risk minimization** (**ERM**). It is *the* central concept in statistical machine learning.

Question: what is the class of function the minimum is taken over?

The class $\{f : \; f$ measurable$\}$ is (way) too large:

- minimization would be computationally intractable,
- we would hopelessly overfit (to be explained in the sequel).

Decision Theory

Let $\mathcal{F}$ be the **hypothesis class** the minimum is taken over:

$$\min_{f \in \mathcal{F}} R_{\mathsf{emp}}(f).$$

Denote any empirical risk minimizer by $\widehat{f}$.

Let us further denote by $\overline{f}$ any minimizer of the risk **over the class** $\mathcal{F}$, i.e.

$$\overline{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} R(f), \qquad \overline{R} := R(\overline{f}).$$

If $\mathcal{F}$ is not further restricted, then $\overline{f} = f_L^*$. The difference between the minimum risk $\overline{R}$ over $\mathcal{F}$ and the minimum risk R^*

$$\mathcal{E}(\overline{f}) := \overline{R} - R^* = R(\overline{f}) - R(f_L^*) = \min_{f \in \mathcal{F}} R(f) - R(f_L^*)$$

is called **excess risk**.

Decision Theory

Relation between empirical risk, excess risk and minimum risk.
Recall

- $\widehat{f} \in \mathcal{F}$ empirical risk minimizer
- $\overline{f} \in \mathcal{F}$ minimizer of the risk over $\mathcal{F}$
- f^*_L minimizer of the risk (without restriction on $\mathcal{F}$)

Theorem

$$\begin{aligned} R(\widehat{f}) &\leq R(\overline{f}) + 2 \sup_{f \in \mathcal{F}} |R_{\mathit{emp}}(f) - R(f)| \\ &= R(f^*_L) + \mathcal{E}(\overline{f}) + 2 \sup_{f \in \mathcal{F}} |R_{\mathit{emp}}(f) - R(f)| \\ &= \textit{intrinsic error} + \textit{approximation error} + \textit{estimation error} \end{aligned}$$

Decision Theory

Proof.

$$
\begin{aligned}
R(\widehat{f}) &= R(\widehat{f}) - R_{\mathsf{emp}}(\widehat{f}) + R_{\mathsf{emp}}(\widehat{f}) - R(\bar{f}) + R(\bar{f}) \\
&\leq R(\widehat{f}) - R_{\mathsf{emp}}(\widehat{f}) + R_{\mathsf{emp}}(\overline{f}) - R(\bar{f}) + R(\bar{f}) \\
&\leq 2 \sup_{f \in \mathcal{F}} |R_{\mathsf{emp}}(f) - R(f)| + R(\bar{f}). \quad \square
\end{aligned}
$$

Decision Theory

Risk of ERM:

$$R(\widehat{f}) \leq R(f_L^*) + \mathcal{E}(\overline{f}) + 2 \sup_{f \in \mathcal{F}} |R_{\mathsf{emp}}(f) - R(f)|$$

$$= \text{intrinsic error} + \text{approximation error} + \text{estimation error}$$

We cannot do anything about the intrinsic error. Even with infinitely many data and unlimited computational power, this term is present.

It turns out that the other two terms are antagonists: reducing one increases the other. Proper choice of the class $\mathcal{F}$ yields a good trade-off between these two terms.
Also known as the **bias-variance trade-off**.

Decision Theory

$$R(\widehat{f}) \leq R(f_L^*) + \mathcal{E}(\overline{f}) + 2 \sup_{f \in \mathcal{F}} |R_{\textsf{emp}}(f) - R(f)|$$

First note: $R(\widehat{f})$ is random variable, because the second term on the r.h.s. is.

In statistical learning theory, one either bounds

$$2 \sup_{f \in \mathcal{F}} |R_{\textsf{emp}}(f) - R(f)|$$

with high probability (i.e. with probability $1 - \delta$, term is bounded by ...) or in expectation.
Deriving such bounds is an art of itself, and involves advanced techniques from the theory of empirical processes.

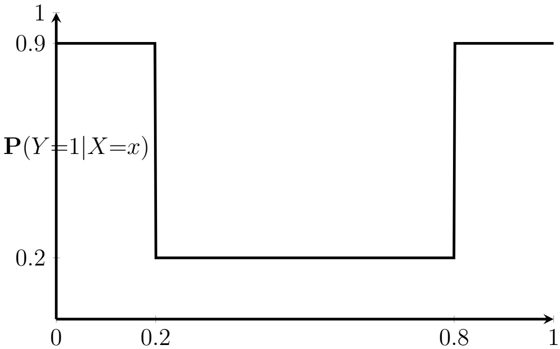
Decision Theory

If $\mathcal{F}$ is "too flexible", then the term

$$\sup_{f \in \mathcal{F}} |R_{\textsf{emp}}(f) - R(f)|$$

becomes large.

Example (c'ted): $\mathcal{X} = [0, 1]$, P_X uniform, $\mathbf{P}(Y = 1|X = x)$ given by



Decision Theory

For the example, given an i.i.d. sample from P of size n, consider

$$\mathcal{F} = \mathcal{F}_n = \mathcal{P}_{n-1} = \left\{ f: \; x \mapsto f(x) = \sum_{k=0}^{n-1} \alpha_k x^k, \; x \in \mathbb{R}, \; \{\alpha_k\}_{k=0}^{n} \subset \mathbb{R} \right\}$$

It is well-known that if $\{(x_i, y_i)\}_{i=1}^{n} \subset \mathbb{R}^2$ have distinct x_i's, there exists (a unique) $p \in \mathcal{P}_{n-1}$ such that

$$p(x_i) = y_i, \quad i = 1, \ldots, n.$$

p is called the interpolating polynomial of $\{(x_i, y_i)\}_{i=1}^{n}$.

Decision Theory

Corollary:

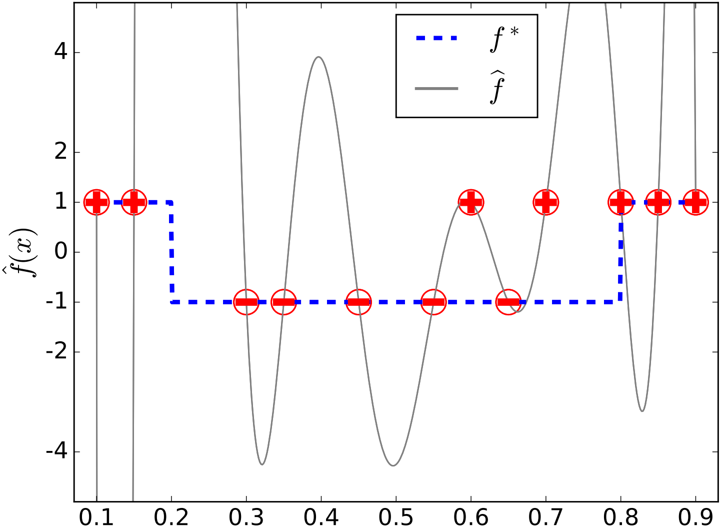
- let $\{(X_i, Y_i)\}_{i=1}^n$ be a sample from P in our example,
- let L be the $0-1$ loss,
- let $\mathcal{F} = \mathcal{P}_{n-1}$ as above.

Then, with probability one,

$$R_{\mathsf{emp}}(\widehat{f}) = \min_{f \in \mathcal{F}} R_{\mathsf{emp}}(f) = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n I(Y_i \neq \mathsf{sign}(f(X_i))) = 0.$$

Decision Theory

Illustration $(n = 12)$:



Decision Theory

While the empirical risk minimizer does well on the observed sample, it **generalizes** poorly to new, yet-to-be-seen data:

Its **generalization error**

$$\mathbf{E}_{(X,Y)\sim P}[L(Y,\widehat{f}(X))|\widehat{f}]$$

is much larger than the empirical risk.

There is no learning process, only memorization of the observed sample (also referred to as **learning set/sample** or synonymously **training set/sample**).

Accordingly, $R_{\mathsf{emp}}(\widehat{f})$ is called **learning/training error**.

The training error can be a poor proxy for the generalization error!

Decision Theory

With an infinite amount of data, we would do well in the example:

There exists a polynomial $\bar{p}$ of degree 2 (i.e. an element of $\mathcal{P}_2$) such that $\text{sign}(\bar{p}) = \text{sign}(f^*)$.

As a result, in the example, the **approximation error**/**excess risk** is zero for $n \geq 2$:

$\min_{f \in \mathcal{F}_n} R(f) = R(f^*) = 0.$

However, because we do not restrict the degree of the polynomial as n grows, the **estimation error** $\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R(f)|$ is out of control. We speak of **overfitting**.

Decision Theory

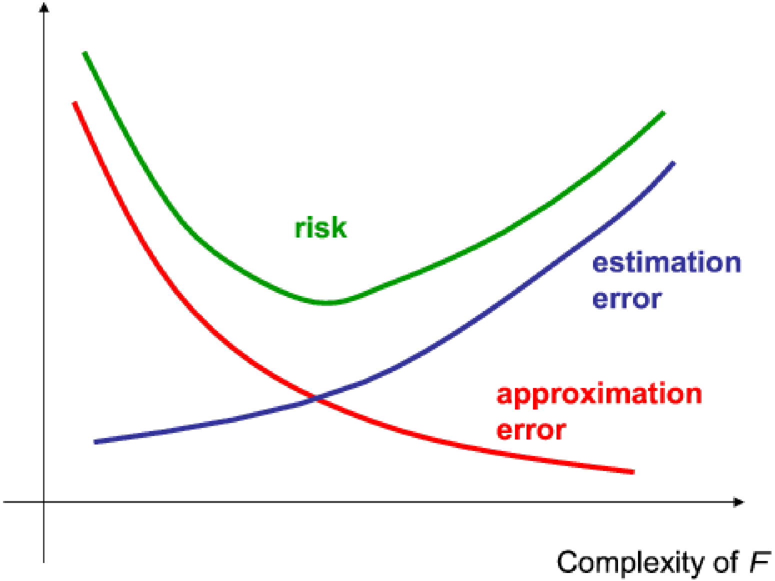
The other extreme is called **underfitting**:

When taking $\mathcal{F} = \mathcal{P}_1$ (affine functions $f(x) = \alpha_0 + \alpha x$), we cannot achieve the Bayes error, no matter how many samples we see. The class $\mathcal{F}$ is not "rich" enough.

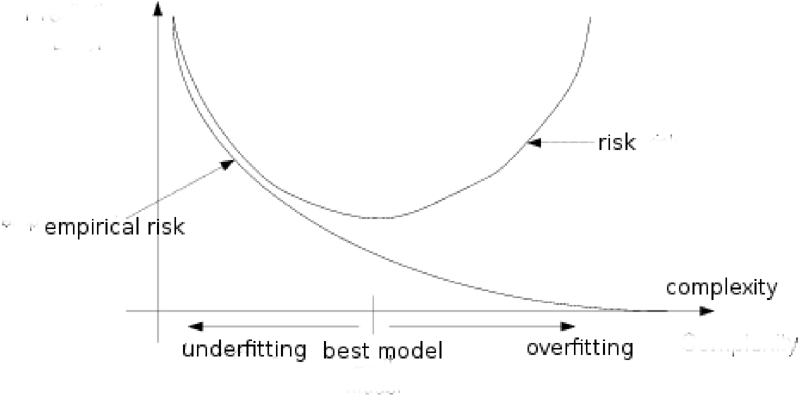
On the other other hand, the estimation error is well-behaved: one can show that with probability at least $1 - \delta$

$$\sup_{f \in \mathcal{F}} |R_{\mathsf{empf}}(f) - R(f)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Decision Theory



Decision Theory



Decision Theory

Some measures of complexity for the class $\mathcal{F}$ (just so you have heard them):

- Rademacher complexity
- Growth function
- VC dimension
- metric entropy
- Gaussian width

These measures are essentially only of theoretical interest.

A principled as well as practical approach to balancing the bias-variance trade-off is **regularization**.